

CERT-PASS

Databricks Databricks Certified Generative AI Engineer

Free Practice Questions Preview

Here are 35 sample questions to help you get started. Unlock the full exam to access all 1100+ questions with detailed explanations.

Question 1 : Design Applications

During a design review, a consumer electronics retailer identifies two constraints: the solution must protect sensitive information, and it must remain maintainable across test and production. The application uses support tickets. Which recommendation is strongest? The design review is focused on: Design a prompt that elicits a specifically formatted response.

- A. Increase temperature so the model explores alternative output formats.
- B. Define the required JSON schema, allowed values, and an instruction to return only schema-valid output.**
- C. Ask the model to be concise but leave the structure implicit.
- D. Add more retrieved documents without defining the response contract.

Answer: B

Explicit structure makes the output contract testable and reduces format drift. The strongest distractor fails because more context can improve factual coverage, but it does not define a machine-readable response contract.

Question 2 : Assembling and Deploying Applications

A university research office is preparing a governed production release for an application using knowledge-base pages. Which option best balances correctness, maintainability, and operational control? The implementation must address this requirement: Preserve memory without unbounded prompt growth.

- A. Keep memory only in the current process.
- B. Duplicate all prior tool results into every prompt.
- C. Append the complete conversation forever.
- D. Store structured state persistently and retrieve a relevant summary or selected facts for each turn.**

Answer: D

Selective persistent memory controls token growth while retaining useful state. The main trap is that unbounded history increases cost and can degrade relevance.

Question 3 : Data Preparation

A university research office has an assistant used by support agents. It processes knowledge-base pages. The current design is causing a problem with retrieval precision. The team also needs to simplify rollback. Which action best addresses the requirement?

A. Add generic payment terms because they also mention orders.

B. Add the authoritative delivery-status source keyed by transaction ID and refresh it as operational data changes.

C. Fine-tune on historic delivery examples so the model predicts current status.

D. Tell the model to add a fixed number of days to every order date.

Answer: B

Operational questions require a fresh authoritative source keyed to the transaction. The main trap is that historical examples cannot reliably answer live shipment-status questions.

Question 4 : Assembling and Deploying Applications

A pilot used by claims reviewers is being moved into production. The application handles equipment logs. Which change best satisfies the requirement?

A. Copy the notebook into a shared folder and use it as the production registry.

B. Log and register the deployable model with MLflow in Unity Catalog so versions and permissions can be governed.

C. Store the model only on a developer laptop.

D. Publish a screenshot of the model configuration.

Answer: B

Unity Catalog registration supports governed lifecycle management. In contrast, the tempting alternative fails because a notebook copy is not a governed model registry.

Question 5 : Design Applications

CASE STUDY: After a release, a automotive parts distributor observes a drop in retrieval recall for an assistant used by support agents. The fix must preserve governance and support controlled rollback. Which action is MOST appropriate? Select the option containing the TWO best actions.

A. Define the required JSON schema, allowed values, and an instruction to return only schema-valid output. Ask the model to be concise but leave the structure implicit.

B. Add more retrieved documents without defining the response contract. Ask the model to be concise but leave the structure implicit.

C. Define the required JSON schema, allowed values, and an instruction to return only schema-valid output. Validate the returned JSON before passing it to downstream systems.

D. Increase temperature so the model explores alternative output formats. Validate the returned JSON before passing it to downstream systems.

Answer: C

Explicit structure makes the output contract testable and reduces format drift. The strongest distractor fails because more context can improve factual coverage, but it does not define a machine-readable response contract. The second selected action is important because validate the returned json before passing it to downstream systems.

Question 6 : Assembling and Deploying Applications

A healthcare operations group is building a Databricks GenAI solution for operations analysts. The workload uses scanned claim forms. Which approach is MOST appropriate? The design review is focused on: Package dependencies for a served RAG chain.

- A. Install packages manually on a developer laptop only.
- B. Declare the runtime dependencies used by the retriever, prompt logic, and model wrapper as part of packaging.**
- C. Assume notebook-installed libraries appear automatically in serving.
- D. Omit retriever dependencies because only the LLM is important.

Answer: B

Served applications need reproducible declared dependencies. The main trap is that notebook state is not a reliable serving environment.

Question 7 : Application Development

CASE STUDY: A retail pharmacy network operates an assistant for support agents. It uses call transcripts. The team must reduce hallucinations, preserve auditability across test and production, and avoid introducing unnecessary latency. Which proposal should be approved? Select the option containing the TWO best actions.

- A. Validate tool arguments against a schema and reject or repair invalid calls before execution. Execute every generated tool call immediately.
- B. Validate tool arguments against a schema and reject or repair invalid calls before execution. Log rejected calls for monitoring and refinement.**
- C. Increase temperature to reduce malformed arguments. Log rejected calls for monitoring and refinement.
- D. Expose internal stack traces directly to end users. Execute every generated tool call immediately.

Answer: B

Schema validation prevents malformed or unsafe calls from reaching downstream systems. The main trap is that immediate execution creates avoidable reliability and security risk. The second selected action is important because log rejected calls for monitoring and refinement.

Question 8 : Assembling and Deploying Applications

An automotive parts distributor compares four implementation proposals for an application that uses customer emails. Which proposal should be selected?

- A. Use metadata filters to generate the embedding vector.
- B. Use embeddings for semantic similarity, an index for retrieval, endpoint capacity for serving, and metadata filters for narrowing results.**
- C. Treat reranking as a replacement for source ingestion.
- D. Use a dashboard as the embedding model and a notebook as the index.

Answer: B

Each Vector Search component has a distinct role. The main trap is that metadata filtering narrows candidates; it does

not create embeddings.

Question 9 : Governance

During a design review, a energy maintenance team identifies two constraints: the solution must support least-privilege access, and it must remain maintainable across development, staging, and production. The application uses customer emails. Which recommendation is strongest?

- A. Ask the model to self-regulate without permission boundaries.
- B. Use one shared unrestricted token for all tools.
- C. Grant each tool only the permissions required for its intended operation and block unauthorized actions.**
- D. Give the agent admin access so it never fails.

Answer: C

Least privilege limits the impact of mistakes or attacks. The strongest distractor fails because admin access increases the blast radius of a prompt-injection failure.

Question 10 : Evaluation and Monitoring

During a design review, a automotive parts distributor identifies two constraints: the solution must support least-privilege access, and it must remain maintainable across development, staging, and production. The application uses scanned claim forms. Which recommendation is strongest?

- A. Sample production traces and score whether the agent selected the appropriate tool or specialist for each request.**
- B. Monitor only CPU utilization.
- C. Count tokens without reviewing routing quality.
- D. Disable trace collection after deployment.

Answer: A

Routing quality requires visibility into the agent's intermediate decisions. The most plausible wrong answer is still incorrect because infrastructure metrics alone cannot show whether the agent selected the correct capability.

Question 11 : Assembling and Deploying Applications

CASE STUDY: After a release, a software-as-a-service company observes a drop in retrieval precision for an assistant used by procurement specialists. The fix must preserve governance and support controlled rollback. Which action is MOST appropriate? Select the option containing the TWO best actions.

- A. Call the embedding model as the response generator. Trace the retriever and model-call steps separately.
- B. Retrieve context before generation and pass the selected evidence into the prompt template. Generate first and retrieve evidence only after users complain.
- C. Retrieve context before generation and pass the selected evidence into the prompt template. Trace the retriever and model-call steps separately.**

D. Write each user's question into the source table before answering. Generate first and retrieve evidence only after users complain.

Answer: C

A basic RAG chain retrieves evidence before the model generates the answer. The main trap is that retrieving afterward cannot ground the original response. The second selected action is important because trace the retriever and model-call steps separately.

Question 12 : Evaluation and Monitoring

CASE STUDY: After a release, a pharmaceutical quality group observes a drop in throughput for an assistant used by procurement specialists. The fix must preserve governance and support controlled rollback. Which action is MOST appropriate? Select the option containing the TWO best actions.

- A. Track live endpoint behavior and investigate regressions using stored requests, responses, and traces. Use only a one-time offline test before launch.
- B. Delete historical inference data before comparing versions. Compare production behavior across versions.
- C. Track live endpoint behavior and investigate regressions using stored requests, responses, and traces. Compare production behavior across versions.**
- D. Monitor only frontend styling performance. Use only a one-time offline test before launch.

Answer: C

Live monitoring is required to detect post-release regressions. The most plausible wrong answer is still incorrect because a one-time test cannot represent changing production traffic. The second selected action is important because compare production behavior across versions.

Question 13 : Assembling and Deploying Applications

A technical lead asks for the best production design for a industrial equipment manufacturer. The workload uses call transcripts, must improve answer grounding, and has a strict audit requirement. Which option best addresses the combined constraints? Select the option containing the TWO best actions.

- A. Use a document OCR tool as the analytics layer. Limit access to the approved data scope.
- B. Grant the agent unrestricted credentials to all databases. Copy every table row into the system prompt.
- C. Expose an approved Genie Space to the supervisor for governed structured-data questions. Copy every table row into the system prompt.
- D. Expose an approved Genie Space to the supervisor for governed structured-data questions. Limit access to the approved data scope.**

Answer: D

A Genie Space supports governed conversational access to structured data. In contrast, the tempting alternative fails because embedding every row in the prompt is inefficient and weakens controls. The second selected action is important because limit access to the approved data scope.

Question 14 : Governance

A Generative AI Engineer is troubleshooting a production application for maintenance engineers. The workflow uses clinical operating procedures, and recent changes caused a regression in answer relevance. Which correction should be made first? The implementation must address this requirement: Use masking techniques as guardrails.

- A. Use a larger model so masking is unnecessary.
- B. Send raw sensitive values and ask the model not to reveal them.
- C. Remove all context even when non-sensitive fields are needed.
- D. Mask sensitive values before they reach the model while retaining the minimum task-relevant context.**

Answer: D

Masking reduces exposure while preserving useful context. The most plausible wrong answer is still incorrect because prompt instructions alone do not reliably prevent sensitive-data leakage.

Question 15 : Data Preparation

During a design review, a construction project office identifies two constraints: the solution must support frequent source updates, and it must remain maintainable across development, staging, and production. The application uses clinical operating procedures. Which recommendation is strongest?

- A. Duplicate each child chunk into multiple indexes without evaluation.
- B. Use one giant chunk per document and remove metadata.
- C. Return isolated sentences and discard their parent relationships.
- D. Use parent-child retrieval so small child chunks support matching while larger parent sections supply coherent context.**

Answer: D

Parent-child retrieval balances precise matching with coherent generation context. The most plausible wrong answer is still incorrect because one giant chunk reduces specificity and can consume excessive context.

Question 16 : Application Development

During a design review, a university research office identifies two constraints: the solution must reduce hallucinations, and it must remain maintainable across sandbox, UAT, and production. The application uses policy manuals. Which recommendation is strongest?

- A. Give the agent broader permissions so it can recover from attacks.
- B. Treat user input and retrieved text as untrusted, preserve instruction hierarchy, and restrict tool permissions.**
- C. Disable logging so malicious payloads are not stored.
- D. Place retrieved content above system instructions.

Answer: B

Layered defenses reduce the impact of prompt injection. The strongest distractor fails because broader permissions increase the blast radius of a successful attack.

Question 17 : Assembling and Deploying Applications

A software-as-a-service company is building a Databricks GenAI solution for operations analysts. The workload uses invoice records. Which approach is MOST appropriate?

- A. Change the model temperature and leave the prompt unchanged.
- B. Overwrite the failed prompt so the old content cannot be recovered.
- C. Delete evaluation history to simplify the registry.
- D. Move the production alias back to the previously approved Prompt Registry version after confirming the regression.**

Answer: D

A versioned registry and aliases make controlled rollback possible. The most plausible wrong answer is still incorrect because overwriting history removes the safest rollback path.

Question 18 : Assembling and Deploying Applications

An automotive parts distributor is preparing a governed production release for an application using knowledge-base pages. Which option best balances correctness, maintainability, and operational control?

- A. Append the entire conversation history forever.
- B. Keep all state only in one process's memory.
- C. Store durable state in a governed datastore and retrieve only the state needed for the current interaction.**
- D. Use browser local storage as the system of record.

Answer: C

Persistent governed storage supports continuity and distributed serving. The strongest distractor fails because process memory disappears across restarts and does not scale reliably.

Question 19 : Application Development

A healthcare operations group has an assistant used by claims reviewers. It processes customer emails. The current design is causing a problem with throughput. The team also needs to reduce hallucinations. Which action best addresses the requirement?

- A. Ignore key fields and retrieve generic FAQs only.
- B. Use temperature to guess missing identifiers.
- C. Append the entire corpus to every request.
- D. Extract the account ID, product family, and intent, then retrieve only the context relevant to those fields.**

Answer: D

Key fields and intent improve retrieval scope and grounding. In contrast, the tempting alternative fails because excess context increases cost and can reduce relevance.

Question 20 : Application Development

A Generative AI Engineer is reviewing a solution for a regional bank. The team works with contracts and needs to reduce hallucinations. What should the engineer recommend? The implementation must address this requirement: Create a prompt that adjusts an LLM response from baseline to desired output.

- A. Add a clear role, constraints, decision rules, examples, and an explicit output contract.
- B. Remove constraints to make the answer more creative.
- C. Add unrelated documents so the prompt appears more comprehensive.
- D. Increase context length without clarifying instructions.

Answer: A

Prompt engineering should state desired behavior directly and testably. The main trap is that a larger context window does not fix ambiguous instructions.

Question 21 : Assembling and Deploying Applications

An automotive parts distributor is preparing a governed production release for an application using HTML help-center pages. Which option best balances correctness, maintainability, and operational control?

- A. Put post-processing only in an unused notebook.
- B. Register only the raw prompt text as the deployable model.
- C. Package preprocessing, model invocation, and output formatting in the pyfunc model so the served contract stays consistent.
- D. Perform preprocessing separately in every client.

Answer: C

A pyfunc wrapper can encapsulate the served behavior consistently. The most plausible wrong answer is still incorrect because client-specific preprocessing creates drift across consumers.

Question 22 : Design Applications

A pilot used by data stewards is being moved into production. The application handles incident reports. Which change best satisfies the requirement? The implementation must address this requirement: Translate business use case goals into desired AI-pipeline inputs and outputs.

- A. Define required user fields, authoritative sources, freshness expectations, output schema, and acceptance criteria.
- B. Build the user interface before identifying source systems and output requirements.
- C. Index every document before deciding which questions the application must answer.
- D. Choose the largest available model first and infer requirements later.

Answer: A

A clear input-output contract drives architecture, data, and evaluation decisions. The main trap is that choosing a model first can waste cost and still leave key data requirements unresolved.

Question 23 : Governance

During a design review, a telecommunications provider identifies two constraints: the solution must protect sensitive information, and it must remain maintainable across sandbox, UAT, and production. The application uses call transcripts. Which recommendation is strongest?

- A. Treat the text as untrusted evidence, keep it below trusted instructions, and prevent it from changing tool policy.
- B. Give the retriever permission to modify access controls.
- C. Execute the document's instructions because retrieval makes them trusted.
- D. Move retrieved text into the system prompt.

Answer: A

Retrieved documents can contain prompt injection and must remain untrusted context. The most plausible wrong answer is still incorrect because retrieval relevance does not confer instruction authority.

Question 24 : Assembling and Deploying Applications

A pharmaceutical quality group is preparing a governed production release for an application using shipment events. Which option best balances correctness, maintainability, and operational control The initial rollout is expected to serve 1008 internal users?

- A. Append the entire conversation history forever.
- B. Use browser local storage as the system of record.
- C. Keep all state only in one process's memory.
- D. Store durable state in a governed datastore and retrieve only the state needed for the current interaction.

Answer: D

Persistent governed storage supports continuity and distributed serving. The most plausible wrong answer is still incorrect because process memory disappears across restarts and does not scale reliably.

Question 25 : Assembling and Deploying Applications

CASE STUDY: After a release, a pharmaceutical quality group observes a drop in answer relevance for an assistant used by claims reviewers. The fix must preserve governance and support controlled rollback. Which action is MOST appropriate? The design review is focused on: Configure Vector Search for a large low-latency workload. Select the option containing the TWO best actions.

- A. Ignore query rate because only embedding count affects performance. Choose a storage-optimized design solely because it sounds cheaper.
- B. Select the index type and retrieval features after benchmarking the required scale, query rate, and latency target. Document the tested capacity assumptions.
- C. Select the index type and retrieval features after benchmarking the required scale, query rate, and latency target. Choose a storage-optimized design solely because it sounds cheaper.
- D. Turn on reranking automatically even when the latency budget is strict. Document the tested capacity assumptions.

Answer: B

Large workloads require workload-specific benchmarking rather than feature guessing. The main trap is that reranking can improve relevance but may add latency, so it must be measured. The second selected action is important because document the tested capacity assumptions.

Question 26 : Governance

A technical lead asks for the best production design for a construction project office. The workload uses support tickets, must reduce maintenance effort, and has a strict audit requirement. Which option best addresses the combined constraints?

- A. Quarantine, redact, replace, or annotate the problematic passage before indexing, then re-evaluate retrieval quality.**
- B. Keep the passage unchanged and hope prompt instructions override it.
- C. Increase overlap so the problematic passage appears in more chunks.
- D. Remove all metadata so the source cannot be traced.

Answer: A

Mitigation should address the source before it repeatedly contaminates retrieval. The most plausible wrong answer is still incorrect because increasing overlap amplifies the problematic content.

Question 27 : Evaluation and Monitoring

A food distribution company is preparing a governed production release for an application using support tickets. Which option best balances correctness, maintainability, and operational control?

- A. Treat model confidence as ground truth.
- B. Use a ground-truth judge when no reference can be defined.
- C. Replace the evaluation set with random prompts.
- D. Use expected-answer or reference-fact judges only when a reliable ground-truth dataset is available.**

Answer: D

Ground-truth-based evaluation requires a dependable reference. The most plausible wrong answer is still incorrect because model confidence is not an external reference.

Question 28 : Application Development

During a design review, a consumer electronics retailer identifies two constraints: the solution must control cost, and it must remain maintainable across test and production. The application uses invoice records. Which recommendation is strongest? The implementation must address this requirement: Compare evaluation and monitoring phases of the GenAI lifecycle.

- A. Treat evaluation and monitoring as interchangeable.
- B. Use curated offline evaluation before release and monitor live traffic and operational signals after release.**
- C. Use monitoring only before release and evaluation only after incidents.
- D. Skip offline evaluation when inference tables are enabled.

Answer: B

Offline evaluation and production monitoring answer complementary questions. The main trap is that inference logging does not replace pre-release testing.

Question 29 : Evaluation and Monitoring

A Generative AI Engineer is troubleshooting a production application for field technicians. The workflow uses contracts, and recent changes caused a regression in retrieval recall. Which correction should be made first? The implementation must address this requirement: Select an LLM choice based on quantitative evaluation metrics.

- A. Choose by parameter count alone.
- B. Choose the model that meets quality thresholds with the best acceptable latency and cost profile on representative tests.**
- C. Choose the cheapest model without checking quality.
- D. Choose the largest model even when a smaller one meets the target.

Answer: B

Model choice should be evidence-based and workload-specific. The most plausible wrong answer is still incorrect because largest-model selection can waste latency and cost budget.

Question 30 : Assembling and Deploying Applications

A pilot used by procurement specialists is being moved into production. The application handles clinical operating procedures. Which change best satisfies the requirement The initial rollout is expected to serve 1230 internal users?

- A. Run one notebook manually for each chat message.
- B. Use `ai_query()` for every chatbot interaction because it is SQL-based.
- C. Use an interactive served agent instead of `ai_query()` when a user needs multi-turn reasoning and tool use in real time.**
- D. Replace tool calls with static batch enrichment.

Answer: C

`ai_query()` fits batch row inference, while interactive multi-turn agents need a served conversational workflow. In contrast, the tempting alternative fails because a batch enrichment function does not provide interactive tool-driven reasoning.

Question 31 : Evaluation and Monitoring

CASE STUDY: After a release, a telecommunications provider observes a drop in cost for an assistant used by research coordinators. The fix must preserve governance and support controlled rollback. Which action is MOST appropriate? Select the option containing the TWO best actions.

- A. Use only an offline benchmark after launch. Monitor only the number of UI visits.
- B. Monitor task quality, groundedness, safety, latency, throughput, error rate, and cost signals relevant to the use case. Monitor only the number of UI visits.

C. Monitor only response length. Set alert thresholds for the most critical metrics.

D. Monitor task quality, groundedness, safety, latency, throughput, error rate, and cost signals relevant to the use case. Set alert thresholds for the most critical metrics.

Answer: D

Production monitoring should combine quality and operational signals. In contrast, the tempting alternative fails because offline benchmarks alone do not reveal live regressions. The second selected action is important because set alert thresholds for the most critical metrics.

Question 32 : Governance

A healthcare operations group has an assistant used by research coordinators. It processes clinical operating procedures. The current design is causing a problem with throughput. The team also needs to simplify rollback. Which action best addresses the requirement?

A. Give the agent admin access so it never fails.

B. Use one shared unrestricted token for all tools.

C. Grant each tool only the permissions required for its intended operation and block unauthorized actions.

D. Ask the model to self-regulate without permission boundaries.

Answer: C

Least privilege limits the impact of mistakes or attacks. The most plausible wrong answer is still incorrect because admin access increases the blast radius of a prompt-injection failure.

Question 33 : Data Preparation

A media subscription service is preparing a governed production release for an application using regulatory PDFs. Which option best balances correctness, maintainability, and operational control?

A. Increase temperature and assume retrieval improved.

B. Measure only final response length.

C. Change the interface layout and compare click rates.

D. Measure retrieval precision and recall against labeled queries before changing the generation model.

Answer: D

Retrieval metrics isolate whether the retriever returns the needed evidence. The most plausible wrong answer is still incorrect because response length does not reveal whether the correct passages were retrieved.

Question 34 : Evaluation and Monitoring

A Generative AI Engineer is troubleshooting a production application for maintenance engineers. The workflow uses support tickets, and recent changes caused a regression in latency. Which correction should be made first?

A. Disable trace collection after deployment.

B. Monitor only CPU utilization.

C. Sample production traces and score whether the agent selected the appropriate tool or specialist for each request.

D. Count tokens without reviewing routing quality.

Answer: C

Routing quality requires visibility into the agent's intermediate decisions. The most plausible wrong answer is still incorrect because infrastructure metrics alone cannot show whether the agent selected the correct capability.

Question 35 : Evaluation and Monitoring

A construction project office is building a Databricks GenAI solution for sales representatives. The workload uses clinical operating procedures. Which approach is MOST appropriate?

A. Use AI Gateway usage tracking, inference tables, and rate limits to govern and observe endpoint consumption.

B. Disable logging and grant unrestricted endpoint access.

C. Put throttling only in browser JavaScript.

D. Rely solely on a monthly cloud bill after usage occurs.

Answer: A

AI Gateway centralizes model-serving governance and observability controls. The most plausible wrong answer is still incorrect because browser-only throttling can be bypassed and is not a robust control.

Unlock All 1100+ Questions

Get the complete Q&A package with detailed explanations, topic analytics, and exam-accurate practice.

From €29.00

Visit: <https://cert-pass.com/exams/databricks-databricks-certified-generative-ai-engineer>

CERT-PASS

cert-pass.com

© 2026 Cert-Pass. This material is for personal use only. Do not distribute.